RESEARCH PLAN PROPOSAL

A new approach for Sentiment Analysis to Measure the Users Opinion by using Machine Learning Techniques

For registration to the degree of
Doctor of Philosophy

IN THE FACULTY OF SCIENCE

IIS UNIVERSITY (DEEMED TO BE UNIVERSITY), JAIPUR

Submitted by

Priyanshu Jadon
Enroll. No.: ICG/2018/26580

Under the Supervision

Dr. Deepshikha Bhatia
Sr. Asst. Professor

Dr. Durgesh Kumar Mishra
Professor (CSE), Director of SAIT Indore

Department of Computer Science & IT

September 2019

1. Topic

A new approach for Sentiment Analysis to measure Users Opinion by using Machine Learning Technique.

2. Introduction:

Sentiment Analysis is a trending topic, as everyone has a unique perception about everything. In today's world every users are using online platform and share their views and opinions and comments regularly. This kind of information can be mined by a various machine learning techniques and used efficiently.

"Sentiment analysis, additionally known as opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities like products, services, organizations, individuals, issues, events, topics, and their attributes." Sentiment analysis refers to the employment of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. Sentiment analysis is widely applied to reviews and social media for a variety of applications, ranging from marketing to customer service.

A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature level—whether the expressed opinion in a document, a sentence or an entity which shows the given feature is positive, negative, or neutral. In advance, sentiment analysis also classifies the sentence on the premise of emotional states such as "happy", "sad", and "angry".

2.1 Classification of Sentiment Analysis:

Sentiment analysis is classified into two techniques: Lexicon based and machine learning based.

In Lexicon based approach works on supporting sentiment counts and weight. With consideration of labeling, on the basis of corpus based, dictionary based and manual approach, viewpoints are integrated. Lexicon-based approach is classified as: Corpus based approach and Dictionary approach.

Machine learning is a category of algorithm that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. The main objective of machine learning is to build algorithm that can receive input data and use statistical analysis to predict an output.

On the basis of machine learning sentimental analysis is classified as:

Unsupervised Learning: From hidden labels, unlabeled data are categorized using unsupervised learning. Although, learner have unlabeled so no error and no potential solution is evaluated.

Supervised Learning: On the basis of common patterns, labeled data are categorized using supervised approach. In testing phase data is classified and data patterns are examined using training data set. These approaches are used to observe and predict data movement. Trained data are analyzed and produces functions for the generation and examining of data class for complete data sample

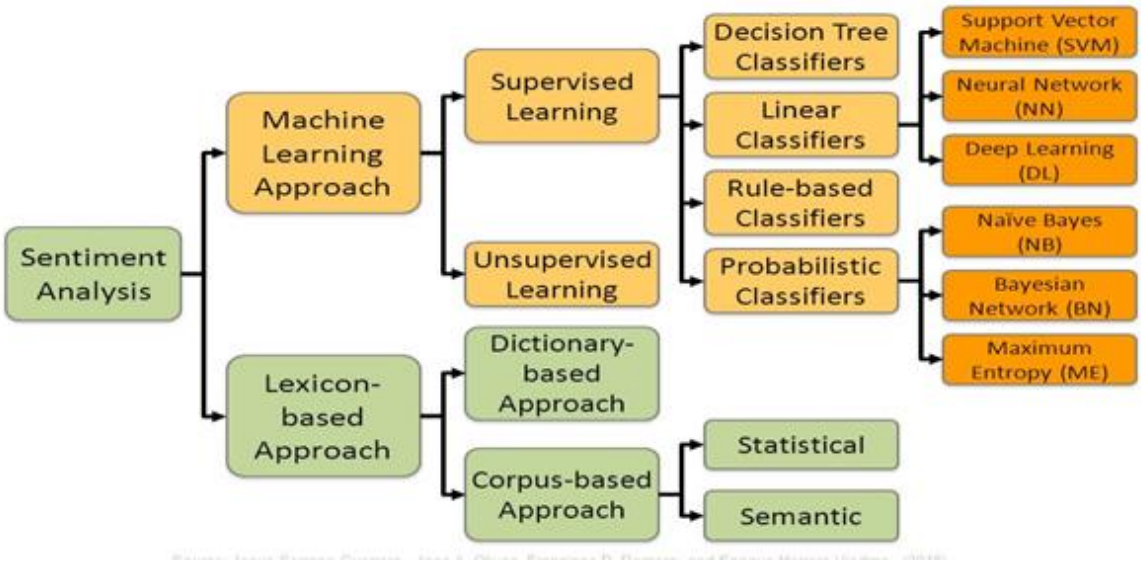

Fig1 : Classification of Sentiment Analysis Technique

2.2 Neural Network:

Neural networks are one of the most beautiful programming paradigm. In the conventional approach to programming, we tell the computer what to do and break big problems up into many small, precisely defined tasks that the computer can easily perform. In contrast, we don't tell the computer how to solve our problems for a neural network. Instead, it learns from observational data and figures out its own solution to the problem. Today, deep neural networks and deep learning achieve outstanding performance for many important problems in computer vision, speech recognition, and natural language processing. They're being deployed on a large scale by companies such as Google, Microsoft, and Facebook.

Neural network are of various types, such as Convolution neural network, Recurrent Neural Network, Recursive neural Network, Feed forward neural network and many more. We are implementing Deep learning with neural network.

The following benefits can be gained by neural network in sentiment analysis process:

- It handles large amount of data sets.
- It has ability to detect complex relationship between variables.
- It achieves high level of learning because it has a huge number of layers.
- It is very beneficial in text generation, sentence classification and feature presentation.

2.3 Cross Validation:

In machine learning, we couldn't fit the model on the training data and can't say that the model will work accurately for the real data. For this, we must assure that our model got the correct patterns from the data, and it is not getting up too much noise. For this purpose, we use the cross-validation technique.

Calculating model accuracy is a critical part of any machine learning project, yet many data science tools make it difficult or impossible to assess the true accuracy of a model. Often tools only validate the model selection itself, not what happens around the selection. Or worse, they don't support tried and true techniques like cross-validation.

2.4 Metrics:

The next step after implementing a machine learning algorithm is to find out how effective is the model based on metric and datasets. Different performance metrics are used to evaluate different Machine Learning Algorithms.

Classification Accuracy: It is what we usually mean, when we use the term accuracy. It is the ratio of number of correct predictions to the total number of input samples.

$$\textbf{ACCURACY} = \frac{\textbf{NUMBER OF CORRECT PREDICTIONS}}{\textbf{TOTAL NUMBER OF PREDICTIONS MADE}}$$

It works well only if there are equal number of samples belonging to each class

Confusion Matrix as the name suggests gives us a matrix as output and describes the complete performance of the model.

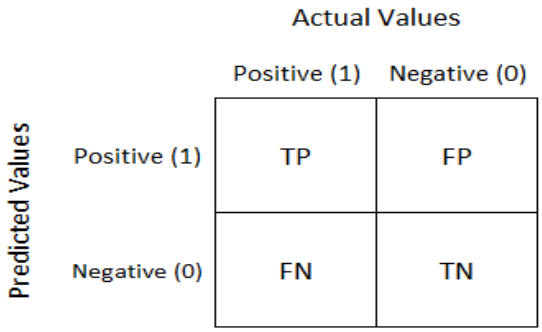

Fig2: Confusion Matrix

There are 4 important terms:

- **True Positives**: The cases in which we predicted YES and the actual output was also YES.
- **True Negatives**: The cases in which we predicted NO and the actual output was NO.
- **False Positives**: The cases in which we predicted YES and the actual output was NO.
- **False Negatives**: The cases in which we predicted NO and the actual output was YES.

Accuracy for the matrix can be calculated by taking average of the values lying across the **"main diagonal"** i.e

$$\textbf{ACCURACY} = \frac{\textbf{TRUE POSITIVES + FALSE NEGATIVES}}{\textbf{TOTAL NUMBER OF SAMPLES}}$$

Confusion Matrix forms the basis for the other types of metrics.

F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. It tells how precise the classifier is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances).

High precision but lower recall, gives us an extremely accurate, but it then misses a large number of instances that are difficult to classify. The greater the F1 Score, the better is the performance of our model. Mathematically, it can be expressed as:

$$\textbf{F1 SCORE} = (2 * \textbf{PRECISION} * \textbf{RECALL}) / (\textbf{PRECISION} + \textbf{RECALL})$$

F1 Score tries to find the balance between precision and recall.

- **Precision:** It is the number of correct positive results divided by the number of positive results predicted by the classifier.

PRECISION = TP/(TP+FP)

- **Recall:** It is the number of correct positive results divided by the number of ***all*** relevant samples (all samples that should have been identified as positive).

RECALL = TP/(TP+FN)

3. Literature Review

Under the Literature Survey, various research papers are studied and that provides the brief knowledge of the research subject to the researcher. Related work based on reviews of author depending upon new trends and emerging technologies which are correlated with each other is present in this part.

In research done by Yang et al (2018) on paper entitled "A survey on Sentiment Analysis by using Machine Learning methods" mainly describes the popular Sentiment Analysis Technique such as : SVM, NB, ME, ANN method. They evaluate the performance on the basis of Accuracy, Precision, Recall and F1 Score metrics. SA analysis methods are in English language. There is a lot of problem if we apply this method to other languages, such as no open source material is available, only paid version is present which is very costly. Some classifier provides high success rate but accuracy percentage is not been achieved.

In research done by Gupta et al (2014) on paper entitled "Comparative Study of classification algorithm used in Sentiment Analysis", this paper provides the study of NB, ME, Boosted Tree, Random Forest classifier. They show the comparison among them and show us that which method gives better accuracy. Main findings are mentioned as: if processing time, memory and less training time are constraints given than use Naïve Bayes classifier. If less training time is available but have high processing time and memory than use Maximum entropy. If our higher

priority is accuracy than use Random Forest. If we want average utilization of CPU resources, accuracy and average time for training the data than use Boosted Tree.

In research done by Trupthi et al (2017**)** on paper entitled "Sentiment Analysis of twitter Data using Streaming API", this paper predicts the sentiment of the tweets of the people posted in social media using hadoop. They are using Naïve Bayes classifier along with uni-word feature extraction. Use Streaming API for fetching tweets and Morris API for applying Filter. They are using uni-word approach so we can't get a proper semantic of the text or word. Hence, try to use N-gram approach for getting a proper semantic of the sentence. For this we have to apply a pattern filters in hadoop. We suggest that use other Machine learning approach rather than NB approach because naïve bayes classifier works on the assumption that one variable is dependent on others.

In research done by Jadav et al (2016) on paper entitled "Sentiment Analysis using Support Vector Machine based on feature Selection and Semantic Analysis", here they proposed a OPTIMIZED SVM classifier. Datasets used here: movie review, twitter, and gold dataset. This datasets are also evaluated on NB and SVM classifier and then the comparison between three has to be done. Accuracy is analysis on the basis of Confusion Matrix. There are many kernel functions present in SVM. Here, Gaussian Radial Basic Kernel was modified. The comparison is made by modifying only the Gaussian Radial Basic Kernel function. We can also modify the other kernel functions and compare it with various other classifiers like SVM, ME, NB, ANN etc.

In research done by Aloui et al (2018) on paper entitled "A novel adaptable approach for sentiment analysis on big social data" provides an adaptable sentiment analysis approach is proposed. Dynamic dictionary of word is constructed here. Classification of tweets is on the basis of 6 different parameters: highly +ve, moderately +ve, lightly +ve, highly –ve, moderately –ve & lightly –ve. Tweets of 2016 US election were taken as a dataset. Comparison is made by evaluating the dataset on Google cloud prediction API and NB. Performance is checked on accuracy, precision, recall, F-measure metrics. The proposed approach is only used on small dataset of twitter. We try to implement the same on large dataset.

In research done by Shaya et al (2018) on paper entitled "Sentiment Analysis of Big Data: Methods, Applications, and Open Challenges" presents a systematic comprehensive literature survey that discuss both technical (based on the techniques used) and non-technical (based on the SA application) aspects. They also show the comparative chart between various techniques such as ANN, RF, SVM, GA, NB, DT, and Ensemble Voted Classifier. Comparison is based on the working principle, advantages and Disadvantages, and time complexity. Various challenges are present: Analyzing Uncertain and Incomplete Data & Heterogeneous characteristics of big data.

In research done by Mohey et al (2016) on paper entitled "Survey on sentiment analysis challenges", mainly presents the SA challenges of different classifiers. Challenges are Negation Handling, Feature extraction, NLP overheads. Second comparison relies on Accuracy rate. Here, Comparison is made on the basis of latest techniques used to analysis Sentiment Analysis.

In research done by Jianqiang et al (2017) under the paper "Comparison Research on Text Pre-processing Methods on text Pre-Processing Methods on Twitter Sentiment Analysis", they are mainly focusing on text pre-processing method of data. Six pre-processing methods they apply: Replacing negative mentions, removing url's link, Reverting words that contain repeated letters, Removing numbers, Removing stop words and Expanding acronyms. They are using a feature extraction model: Word n-gram features model and Prior Probability Score feature model. Classifiers used in work are: Support Vector Machine, Naïve Bayes, Logistic Regression, and Random Forest. Scikit-learn library is also used for the classification of data. Classification task performed: a Binary task, and 3-way Task. Data performance has been evaluated on accuracy and F-measure parameters. Data sets used in the method are Stanford Twitter Sentiment Test (STS-test), SemEval 2014, Stanford Twitter Sentiment Gold (STS-gold), Sentiment Evaluation (SE-Twitter). Removing URL's barely affects the performance of classifier in the two feature model. URL's do not contain useful information. Removing stop words has few effects on the performance of classifier in n-gram model while prior polarity model leads to the fluctuation of classifier performance. Removing numbers has no effect on the accuracy of sentiment classification in the prior polarity model but in n-gram model it causes fluctuation o SVM classifier. We say that removing word in SVM classifier improves its accuracy. Removing repeated letters on the performance of classifier is different on each dataset. Expanding Acronym improves the performance of classifiers on most datasets.

In research done by Park et al (2018) under the paper entitled "Sentiment Analysis of Twitter Corpus Related to Artificial Intelligence Assistants", uses a lexicon based sentiment analysis where they used **V**alence **A**ware **D**ictionary and s**E**ntiment **R**easoner (VADER) to classify the tweets as a positve, negative and neutral. The user's opinion about three AI assistants: Siri by Apple, Google Assistant by Google, and Cortana by Microsoft are examined. They perform the t-Test, Kruskal-Wallis test, and Mann-Whitney test to show the statistical significance. Data is fetched by using twitter Streaming API. The Google Assistant had the highest rank and Siri had the lowest rank. The author established a criterion, which could make a decision to choose a better AI assistant of the tweets. They are only using lexicon based approach we try to do same thing on a machine learning approach.

In research done by Jianqiang et al (2018) under the paper entitled "Deep Convolution Neural network for Twitter Sentiment analysis*",* here the author apply a convolution algorithm on twitter sentiment analysis to train deep neural network, in order to improve accuracy and analysis speed. Feature extraction method used: Word n-gram (uni and bi) feature model. The feature set is combined and fed into deep convolution neural network to train and predict the tweet as a positive or negative. The author proposed a GloVe-DCNN model which implements the binary task to classify the tweets. Datasets as paper 8 and classifier are SVM and LR with n-gram, SVM and LR with GloVe, and GloVe-DCNN. Evaluation of performance is done on Recall, Precision and F-measure. 87.62% is a higher accuracy which was achieved by DCNN on Stanford Twitter Sentiment Test (STS-T). The DCNN gives a better performance than other classifiers. It avoids the error occurrence and improves the classification performance.

In research done by Bouazizi et al (2017) under the paper entitled "A Pattern-Based approach for Multi-class Sentiment Analysis in Twitter", mainly proposed a novel approach which classifies the text into multiple sentiment classes. They are using 7 different sentiment classes, like Fun, Happiness, Love, Neutral, Sadness, Anger and Hate**.** They introduce a SENTA tool where everything is going to happen. The user doesn't need an extra tool or library for data pre-processing or feature extraction. Everything is done on a SENT. They take a 2 different data set of twitter and perform a binary, 3-way and multi-class classification on it. The comparison analysis is based on accuracy, precision, recall and F-measure metrics.

In research done by Chen et al (2018) under the paper entitled "Deep Convolution Network for twitter Sentiment Analysis", introduces word embedding feature extraction method and combine this is combined with n-gram feature and word sentiment polarity score feature and form new sentiment feature set of tweets. The feature set of tweet is integrated into the deep convolution neural network for training and predict the sentiment of the user behind the textual data.

In research done by Chen et al (2018**)** under the paper entitled "A Deep Neural Network for target based Sentiment Analysis", proposes a deep learning neural network model combining Convolution Neural Network and Regional Long Short Term Memory (RLSTM). RLSTM is used to reduce the training time. This approach is used in aspect based sentiment analysis which infers the sentiment polarities of different aspects in the same sentences.

In research done by Jain et al (2018) under the paper entitled "A comparative study of machine learning and Deep learning techniques for Sentiment analysis", gives the comparison between various machine learning as well as their hybrid techniques and get that deep learning techniques will gives a better result.

In research done by Paliwal et al (2018) under the paper entitled "Sentiment Analysis and Prediction using Neural Network", talks about how ANN performs Sentiment analysis. They are using Feed Forward Neural Network to test the data and Back Propagation Neural Network is used to train the data.

In research done by Dhar et al (2018) under the paper entitled "A sentiment Analysis using neural network: A new approach", uses a CNN with new Morphological Sentence Pattern model.

Above study shows that very little work is going to be done on vector space feature extraction methods. Feature extraction is divided into two parts in Neural Network: Meta Feature and Vector Space Feature. Everyone is going to extract the Meta features. We mainly emphasis on Vector space features where count vectorization, TF-IDF Vectorizer, Word Embedding, Hash vectorizer. In above we also find that some are using Word embedding vectorizer. In our research we are going to use every TF-IDF Vectorizer, Word Embedding and hash vectorizer feature extraction method. We also focus on dimension reducing for achieving better performance. For Classification task we are proposing a feed forward and back propagation neural network algorithm which helps us to train, validate and test data. Before entering into the NN we pre-train our data using Naïve bayes with n-gram approach and then send it to the input

layer of NN for achieving better performance. To reduce the training time in NN we are using a LSTM (Least short time memory) technique.

In above study they classify the tweets in three categories: +ve, -ve, neutral but we are going to classify the text into 5 categories: more +ve, less +ve, neutral, more negative, less negative

3. Motivation:

According to Ramteke et al. (2012) motivation for Sentiment Analysis is two-fold. Both consumers and producers highly value "customer's opinion" about products and services. Thus, Sentiment Analysis has seen a considerable effort from industry as well as academia.

Hate Speech is an unfortunately common occurrence on the internet often social media sites like FB & Twitter face the problem of identifying problematic post. While taking decision it is important to know the opinion of the people. It is very difficult to extract the information from the large dataset. So we need a system which differentiates between good reviews and bad reviews. So, through sentiment Analysis we can find out user's opinion in terms of positive and negative perspective. Through this vendors find out the customers views on particular product or service which helps him to increase the business in future.

5. Objectives:

The major objectives set for the research work are:

1) To study and review the existing techniques of machine learning and identifying datasets to be used.

2) Preprocessing of dataset.

3) To improve the feature extraction phase of sentiment analysis.

4) Implementation of algorithm.

5) To validate the data set before it is going to be tested.

6) To predict the text based data into different categories

7) To analyze the result by using performance metrics.

6. Plan of Work and Methodology

6.1 Work Plan:

Phase 1 (4-5 months):

Literature survey on Neural Network, Cross validation of data, and how to fetch the data through various API's available in python environment. (Learn how to implement the libraries and APIs in Python)

Phase 2 (2-3 months)

Detail study of how pre-processing of datasets is going to be done and feature extraction in neural network is implemented.

Phase 3 (6-7 months)

Detail study of various algorithms, to combine various feature extraction methods into new feature dataset and making a Neural Network model.

Phase 4 (4-10 months):

Detail study of integrating a new featured data into NN and applying the NN model for training, validation and testing and getting the result into different categories.

Phase 5 (4-5 months):

Analyze and plot the result obtained by checking the performance by creating the confusion matrix and measure the performance by calculating the computational accuracy with the help of Recall, precision and F-measure and thesis writing.

6.2 Methodology:

To attain the above mentioned objective, the following methodology will be adopted:

- Fetching the data by using the API's available.
- After fetching the data, firstly the Data Pre-processing is done, where we are doing the following: converting the data into lowercase, Drop Non alphabets, Drop URLs, Drop Stop words, Lemmatization of words, Drop unwanted words to remove noise, and Parts of Speech (POS) tagging. Secondly, feature extraction is done where we extract the most important sentiment features through various feature models.
- Now we combine all above feature set model by using any machine learning algorithm (let's say by combining SVM and Linear Regression) .

- New feature set is present which is feed into the Neural network where our data is trained by Back Propagation neural network validate by 10-fold cross validation method, and test by feed forward neural network. Validation is done as it ensures that our data is not over fitting.

- Now, we predict the output (with the help of test data) as less positive, most positive, neutral, less negative & more negative.

- At last we are checking the performance of our result with the help of computational accuracy, Recall, Precision, F1 Score.

- Conclusions will be drawn on the basis of the above experimentation. Lastly discussion will be done leading to final conclusions and scope for future work.

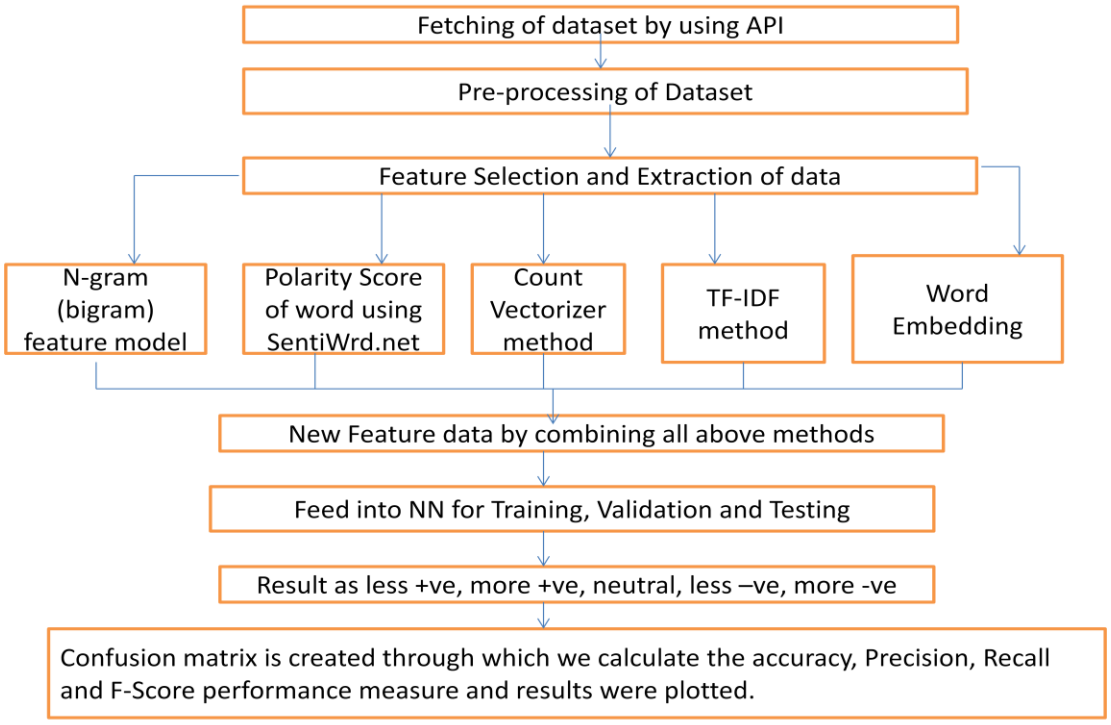
Fetching of dataset by using API
Pre-processing of Dataset
Feature Selection and Extraction of data
N-gram (bigram) feature model
Polarity Score of word using SentiWrd.net
Count Vectorizer method
TF-IDF method
Word Embedding
New Feature data by combining all above methods
Feed into NN for Training, Validation and Testing
Result as less +ve, more +ve, neutral, less –ve, more -ve
Confusion matrix is created through which we calculate the accuracy, Precision, Recall and F-Score performance measure and results were plotted.

Fig: Flowchart depicting methodology adopted

7. Place of work and facilities available

To follow the line of investigation the research will be carried out in the laboratories existing in the university and the workplace where I am doing job. The hardware and software available here will be used as a part of the research work.

8. References, bibliography and webliography

1] Gupta A., S. Joshi, P. Gadgul, and A. Kadam (2014). Comparative Study of classification algorithm used in sentiment analysis. *International Journal of Computer Science and Information Technologies.* **5(5),** 6261-6264

[2] Jadav B., V. Kumar, and B. Vaghela (July 2016). Sentiment Analysis using Support Vector Machine based on feature Selection and Semantic Analysis. *International Journal of Computer Application*. **146-No. 13**, 26-30

[3] Mohey D., and M. Hussain (2016). A survey on sentiment analysis challenges. *Journal of King Saud University- Engineering Science. 1-9*

[4] M. Trupthi, S. Pabboju, and S. Narasimha (2017) .Sentiment Analysis of twitter Data using Streaming API. *IEEE 7th International Advance Computing Conference*. 915-919

[5] Zha J. , and G. Xiaolin (2017). Comparison Research on Text Pre-processing Methods on text Pre-Processing Methods on Twitter Sentiment Analysis. *IEEE, Translations and Content Mining. vol 5,2870-2879*

[6] Yang P. , and Y. Chen (2018). A survey on Sentiment Analysis by using Machine Learning methods. *IEEE 2nd Information Technology, Networking Electronic and Automation Control Conference ITNEC*. 117-121

[7] Aloui I. , Y. Gahi, R. Messoussi, Y. Chabbi, and A. Alexi (2018). A novel adaptable approach for sentiment analysis on big social data. *Journal of Big Data, Springer Open.* 3-18

[8] Shaya S., N. Ismawati, S. Bahri, A. Sulaiman, P. Seuk Wai, and Y. Wai Chung (2018). Sentiment Analysis of Big Data: Methods, Applications, and Open Challenges. *IEEE Access, Magazine and Journal*. **6**, 37807-37827.

[9]Park Won C. , and D. R. Seo (2018).Sentiment Analysis of Twitter Corpus Related to Artificial Intelligence Assistants. *5th International Conference on Industrial Engineering and application.495-498*

[10]Jianqiang Z., G. Xiaolin, and Xuejun Z (2018).Deep Convolution Neural network for Twitter Sentiment analysis. *IEEE, Translation and Content Mining*.23253-23259

[11] Bouazizi M., and T. Ohtsuki (2017). A Pattern-Based Approach for Multi-Class Sentiment Analysis in Twitter. *IEEE-Translations and Content Mining.* 20617-2326.

[12]Chen S. , and P. Chao (2017). Deep Convolution Network for twitter Sentiment Analysis. *IEEE-Translations and Content Mining.* 2169-3536.

[13] Chen S., P. Chao, L. Cai, and L. Guo(2018). A Deep Neural Network for target based Sentiment Analysis. *IEEE- International Joint Conference on Neural Networks.*

[14] Jain K., and S. Kaushal (2018). A comparative study of machine learning and Deep learning techniques for Sentiment analysis". *7th International Conference on Reliability, Infocom Technologies and Optimization(ICRITO).483-487.*

[15] Paliwal S., SK. Khatri Sunil, and M. Sharma (2018). Sentiment Analysis and Prediction using Neural Networks. *International Conference on Inventive Research in Computing Applications (ICIRCA).* 1035-1042.

[16] Dhar S., S. Pednekar, K. Borad, and Prof. A. Save (2018). A sentiment Analysis using neural network: A new approach. *2nd International Conference on Inventive Communication and Computational Technologies (ICICCT).*1220-1224.